

# Краткий обзор генома и протеома бактерии *Aeromonas hydrophila* YL17

Антоненкова Ю. С.

Факультет биоинженерии и биоинформатики МГУ имени М.В. Ломоносова

## РЕЗЮМЕ

Данный обзор представляет собой анализ генома и протеома бактерии *Aeromonas hydrophila* YL17. В работе рассмотрены следующие их характеристики: деление генов по категориям, распределение генов по двум цепям ДНК, составление гистограммы длин белков.

## 1 ВВЕДЕНИЕ

*Aeromonas hydrophila* (гидрофильная аэромонада) — гетеротрофная грамтрицательная палочковидная бактерия, обитающая в основном в районах с тёплым климатом и в пресной, и в солёной воде. Выживает как в аэробной, так и в анаэробной среде и может разлагать желатин и гемоглобин. Устойчива к большинству распространенных антибиотиков и низким температурам. Передвигается с помощью полярного жгутика, имеет капсулу. Может расти при низких температурах вплоть до 4 °С. *A. hydrophila* считается условно-патогенным микроорганизмом. У человека она может вызвать гастроэнтерит [1]. Геном представлен кольцевой ДНК (идентификатор GeneBank: CP007518.2), размер генома составляет 4796281 пар нуклеотидов, содержание GC пар - 61.60% [2].

В ходе работы были рассмотрены геном и протеом бактерии. Задачи работы – определить количество генов разных категорий, визуализировать количество белков разных длин, а также посчитать распределение генов по цепочкам ДНК.

## 2 МАТЕРИАЛЫ И МЕТОДЫ

При написании обзора использовались данные о генах с сайта NCBI (главным образом, файл GCA\_000612075.2\_ASM61207v2\_feature\_table.txt.gz из базы данных NCBI [3]). Обработка данных произведена в программе Microsoft Excel 2013.

Для начала данные из вышеупомянутого файла были импортированы в таблицу Excel, текст был разбит по столбцам.

Для подсчета количества генов разных типов была создана сводная таблица. Гены были распределены по категориям согласно их обозначениям в исходном файле, которые представлены в таблице 1.

Далее были проанализированы длины белков. Значения длин были взяты из колонки product\_length. Белки были распределены по группам в зависимости от их длины, была построена гистограмма, представляющая соотношение количеств белков разных длин. Интервал длин по умолчанию 100 аминокислотных остатков, то есть группы белков следующие: 1-99 а. о., 100-199 а. о., 200 – 299 а. о. и т. д. В таблице Excel из сопроводительных материалов интервал можно изменить вручную, в

этом случае гистограмма обновится (необходимо заметить, что при значении интервала менее 25 и более 150 гистограмма не репрезентативна).

С помощью встроенных функций были рассчитаны некоторые статистические данные. В таблице 2 можно ознакомиться с примененными функциями и их значением в данной работе.

Категория гена	#feature	Class
Кодирующий белок	gene / CDS	protein_coding / with_protein
Псевдоген	gene	pseudogene
тРНК	gene	tRNA
рРНК	gene	rRNA
Не кодирующая РНК	gene	RNase_P_RNA

Таблица 1. Соответствие категорий генов и их обозначений.

Название функции	Цель
СЧЁТЕСЛИМН	Количество белков, длина которых попадает в заданный диапазон; распределение белков по группам в соответствии с их длиной
МИН, МАКС	Минимальная и максимальная длина белка соответственно
СРЗНАЧ	Средняя длина белков
СТАНДОТКЛОН	Среднее квадратичное отклонение от среднего, оно же стандартное отклонение
МЕДИАНА	Медиана длин белков

Таблица 2. Использование функций Excel.

Далее было проанализировано распределение генов по цепям ДНК. Для этого использовался столбец strand исходной таблицы. Значение «+» соответствует прямой цепи, «-» - обратной. Для подсчета использовалась функция СЧЁТЕСЛИМН.

## 3 РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

### 3.1 Количество генов разных категорий

Категория	Количество генов
Кодирующие белок	4078
Псевдоген	35
тРНК	125
рРНК	31
нкРНК	1
Всего	4270

Таблица 3. Количество генов по категориям

Из таблицы 3 видно, что больше всего – 4078 – генов, кодирующих белки. В категориях генов, кодирующих РНК, большинство составляют гены тРНК, а ген не кодирующей РНК в геноме всего один. Из таблицы 4 видно, что на 1 млн п. н. приходится примерно 890 генов.

Количество генов	Длина генома, п. н.	Генов на 1 млн п. н.
4270	4796281	890,3

Таблица 4. Количество генов на 1 млн п. н.

### 3.2 Распределение длин белков

Был проведен анализ длин белков, были получены статистические данные, описанные в таблице 5.

Min длина	25
Max длина	9603
Средняя длина	332,5
Станд. отклонение	270,8
Медиана	286

Таблица 5. Статистика длин белков.

Из таблицы видно, что, несмотря на большой разброс значений длин белков (от 25 до 9603), средняя длина и медиана относительно малы и примерно равны 300. Также была получена гистограмма (рисунок 1). Из нее видно, что большинство белков (около 77%) имеют длину от 100 до 499, что согласуется со статистическими данными.

### 3.3 Распределение генов по цепям ДНК

Подсчет количества генов на прямой и обратной цепях ДНК не показал выраженных особенностей их распределения. В таблице результатов (таблица 6) можно увидеть несильные расхождения между значениями: гены белков и псевдогены в чуть большем количестве находятся на прямой цепи, гены РНК, напротив, на обратной.

Цепь\Тип гена	Гены белков	Псевдогены	Гены РНК
Прямая	2085	19	68
Обратная	1993	16	89

Таблица 6. Распределение генов по цепям ДНК



Рисунок 1. Гистограмма распределения белков по группам в зависимости от их длины.

## СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

[http://kodomofbb.msu.ru/~yulia\\_ant/term1/pr13\\_Antonenkova.xls](http://kodomofbb.msu.ru/~yulia_ant/term1/pr13_Antonenkova.xls)  
x - таблица Excel

## БЛАГОДАРНОСТЬ

Хотелось бы поблагодарить коллектив преподавателей биоинформатики за полученные знания и интересные задания.

## СПИСОК ЛИТЕРАТУРЫ

- [1] [https://ru.wikipedia.org/wiki/Aeromonas\\_hydrophila](https://ru.wikipedia.org/wiki/Aeromonas_hydrophila) - общая информация о бактерии
- [2] <https://www.ncbi.nlm.nih.gov/genome/genomes/1422> - здесь можно ознакомиться с характеристиками *Aeromonas hydrophila* YL17 и сравнить их с характеристиками других штаммов *A. Hydrophila*
- [3] [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/612/075/GCA\\_000612075.2\\_ASM61207v2](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/612/075/GCA_000612075.2_ASM61207v2) - директория NCBI с нужным файлом